A novel real-time, lightweight chaotic-encryption scheme for next-generation audio-visual hearing aids

Adeel, Ahsan ; Ahmad, Jawad ; Larijani, Hadi; Hussain, Amir

Published in:
Cognitive Computation

DOI:
10.1007/s12559-019-09653-z

Publication date:
2020

Document Version
Author accepted manuscript

A novel real-time, lightweight chaotic-encryption scheme for next-generation audio-visual hearing-aids

Ahsan Adeel*[1,2], Jawad Ahmad[3], Hadi Larijani[3], and Amir Hussain[4]

[1]University of Wolverhampton, Wolverhampton, WV1 1LY, UK
[2]deepCI.org, 20/1 Parkside Terrace, Edinburgh, EH16 5XW, UK
[3]Glasgow Caledonian University, Glasgow, G4 0BA, UK
[4]Edinburgh Napier University, Edinburgh, EH10 5DT, UK
[1,2]a.adeel@wlv.ac.uk; ahsan.adeel@deepci.org; a.hussain@napier.ac.uk

November 30, 2019

Structured Abstract

Objective:

Next-generation audiovisual (AV) hearing-aids stand as a major enabler to realise more intelligible audio. However, high data rate, low latency, low computational complexity, and privacy are some of the major bottlenecks to the successful deployment of such advanced hearing-aids. To address these challenges, we propose a novel framework based on an integration of 5G Cloud-Radio Access Network (C-RAN), Internet of Things (IoT), and strong privacy algorithms to fully benefit from the possibilities these technologies have to offer.

Background:

Existing audio-only hearing-aids are known to perform poorly in noisy situations where overwhelming noise is present. Current devices make the signal more audible but remain deficient in restoring intelligibility. Thus, there is a need for hearing aids that can selectively amplify the attended talker or filter out acoustic clutter.

Methods:

The proposed 5G IoT enabled AV hearing-aid framework transmits the encrypted compressed AV information and receives encrypted enhanced reconstructed speech in real-time to address cybersecurity attacks such as location privacy and eavesdropping. For security implementation, a real-time lightweight AV encryption is proposed, based on a piece-wise linear chaotic map (PWLSM), Chebyshev map, and a secure hash and S-Box algorithm. For speech enhancement, the received secure AV (including lip-reading) information in the cloud is used to filter noisy audio using both deep learning and analytical acoustic modelling. To offload the computational complexity and real-time optimization issues, the framework runs deep learning and big data optimization processes in the background, on the cloud.

Results:

The effectiveness and security of our proposed 5G-IoT-enabled AV hearing-aid framework are extensively evaluated using widely known security metrics. Our newly reported, deep learning-driven lip-reading approach for speech enhancement is evaluated under four different dynamic real-world scenarios (cafe, street, public transport, pedestrian area) using benchmark Grid and ChiME3 corpora. Comparative critical analysis in terms of both speech enhancement and AV encryption demonstrate the potential of our envisioned technology to deliver high quality speech reconstruction and secure mobile AV hearing aid communication.

Conclusion:

We believe our proposed 5G IoT enabled AV hearing aid is an effective and feasible solution and represents a step change in the development of next generation multimodal digital hearing aids. The ongoing and future work includes more extensive evaluation and comparison with benchmark lightweight encryption algorithms and hardware prototype implementation.

Keywords– Hearing Aid, 5G Cloud-Radio Access Network, Internet of Things, Cybersecurity, Speech Enhancement, Deep Learning

1 Introduction

Hearing impairment is a hidden disability with no painful symptoms. People with serious hearing-loss find themselves socially isolated and depressed with more negative consequences including headaches, muscle tension, increased stress, insecurity, and sadness [1]. Hearing aids are the most widely used devices for the majority of hearing impairments. The global hearing aid industry, estimated at around US $6.97 billion in 2017, is expected to grow at 7 percent (compound annual growth rate) by 2022, reaching USD 9.78 Billion, according to the market research firm MarketsandMarkets [2]. However, existing hearing aids often perform poorly for speech in noise. Current devices make the signal more audible but remain deficient in restoring intelligibility i.e., no improvement in Signal-to-noise ratio (SNR). Thus, the existing audio-only hearing-aids are not robust to reverberation; therefore intelligibility wins at the cost of higher cognitive load in a noisy environment [3][4].

In recent literature, extensive research has been carried out to develop robust speech enhancement frameworks [5][6][7] [8][9]. However, only a few speech enhancement algorithms have been shown to reliably increase the intelligibility of speech in noise, especially in extreme noisy conditions such as a cocktail party. A limited number of research developments in this field have been implemented into commercially available hearing-aids. For example, spectral subtraction can be very effective in stationary conditions, but the processed speech remains unintelligible. In case of multiple microphones availability, beamforming algorithms could potentially lead to improvements in speech intelligibility. However, such approaches are difficult to employ in unpredictable noisy situations. Recent advances have enabled high data rate and low-latency wireless solutions, which have primarily reformed the innovation direction of the hearing industry. Nevertheless, even sophisticated commercial HAs e.g., latest low-latency and low-cost Bluetooth-enabled HAs, are based on audio-only processing, which remain ineffective in noisy situations. Consequently, existing audio-only hearing aid approaches achieve benefit by simply amplifying the signal, which offers little benefit for understanding speech in high levels of noise [10].

Human performance in noisy environments is known to be dependent on both aural and visual cues, which are combined by sophisticated multi-level integration strategies to improve intelligibility. The multimodal nature of speech is well established in literature, and it is well understood how speech is produced by the vibration of vocal folds and the configuration of articulatory organs. The correlation between the visible properties of articulatory organs (e.g., lips, teeth, tongue) and speech reception has been previously shown in numerous behavioural studies [11][12][13][14]. Therefore, clear visibility of some articulatory organs could be effectively utilized to extract a clean speech signal out of a noisy audio background. The main advantage of using visual cues to extract clean audio

features is their inherent noise immunity [15].

Nevertheless, embracing the multimodal nature of speech presents both opportunities and challenges for hearing assitive technology. The real-time implementation of AV hearing-aids demands high datarate, low latency, low computational complexity, and high security. To address these requirements, IoT stands as a major enabler. However, the growth of IoT raises new radio resource management (RRM) challenges in resource constrained wireless communication systems. Existing wireless systems remain deficient in complying with the huge connectivity requirements of IoT. In contrast, 5G wireless networks address this limitation by exploiting emerging wireless technologies including mmWave, massive MIMO, and C-RAN [16]. In addition, researchers have recently proposed several potential data delivery approaches. For example, the authors in [17] present an overview and highlight the importance of 5G small cell technology in providing high data rate and further improving coverage and capacity in a cost effective manner. Other recently proposed relevant approaches include [18][19][20]. Similarly, researchers have proposed new approaches to address security issues such as [21] where authors proposed a confidential smart-sensing framework in the IoT era with authentication, confidentiality and integrity features.

In this paper, we propose a novel integration of AV speech enhancement technology, 5G, IoT, and strong privacy algorithms. The AV speech enhancement technology is comprehensively presented in our previous works [22][23]. For communication, 5G C-RAN [16] is proposed, which is a widely accepted IoT solution for high data rate, coverage, capacity, and energy efficiency [16]. For security, the lightweight chaotic encryption is proposed and evaluated in this paper. The proposed 5G IoT enabled AV hearing-aid framework is envisioned to address challenges such as cybersecurity attacks (location privacy, eavesdropping), interference between medical IoT devices (that can cause hearing-aids to operate incorrectly with potentially life threatening consequences), low-cost wireless technology design, low power consumption, limited battery, and high datarate requirement. Inspired by our previous work [24], the novel wireless hearing-aid framework offloads computational complexity and real-time optimization issues by running deep learning and big data optimization algorithms in the background on the cloud. The hearing-aid transmits the encrypted compressed audio/visual information to the cloud and receives encrypted enhanced reconstructed speech in real-time. The hearing-aid connects to an indoor 5G wireless access point and back/fronthaul core network that serves as the communication infrastructure of the system.

The rest of the paper is organized as follows: Section 2 presents the proposed 5G IoT enabled AV-hearing aid framework. Section 3 presents the proposed real-time lightweight chaotic encryption algorithm. Section 4 explains the proposed speech enhancement framework including the designed enhanced visually derived Wiener filter (EVWF) and long-short term memory (LSTM) based lip-reading regression model. In Section 5, the used AV datasets and feature extraction methodologies are presented. Section 6 presents the performance evaluation of the proposed AV encryption and speech enhancement algorithms. Finally,

Section 6 concludes this work.

2 5G IoT enabled Audio-Visual Hearing-Aid Framework

Modern digital hearing aids are marvels of sophisticated engineering. To hear modern audio, a low-latency and high datarate wireless solution is needed, that would enable in-ear hearing devices to connect seamlessly [25]. In this article, a novel 5G IoT enabled audiovisual hearing-aid framework is proposed to acquire desired high quality processed speech in noisy environments. An example of state-of-the-art 5G IoT architecture is shown in Figure 1. The IoT enabled devices, supporting a wide variety of applications, to connect to the Internet, whilst utilizing gateway for connectivity. For gateway design, different access technologies such as WiFi and 4G LTE could be used. However, they are both incapable of supporting thousands of connected IoT devices. WiFi suffers from packet collision and limited quality of service (QoS), whereas 4G LTE suffers from high delay and high packet loss for large number of users [16]. In addition, the wireless systems operating in unlicensed frequency bands require additional network equipment, resulting in extra operation and capital expenditures. The unlicensed solutions are also prone to congestion with an exponential increase in IoT deployment. In contrast, next generation 5G wireless networks [26][27] are capable of providing higher datarates, enhanced mobile coverage, improved user experience at relatively lower cost and dense connectivity [28]. Furthermore, to address aggravating detrimental greenhouse (CO2) gas emissions due to ultra-dense 5G wireless networks and increased network's energy consumption, 5G C-RAN is a widely accepted solution that enables improved environmental sustainability, OPEX, resource management, and energy efficiency [16].

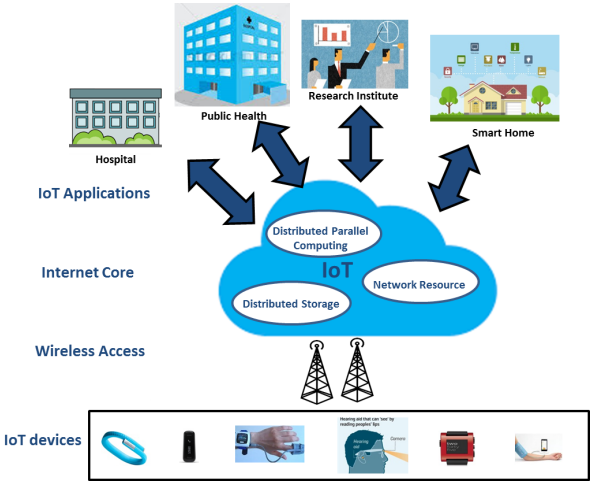


Figure 1: 5G IoT Architecture

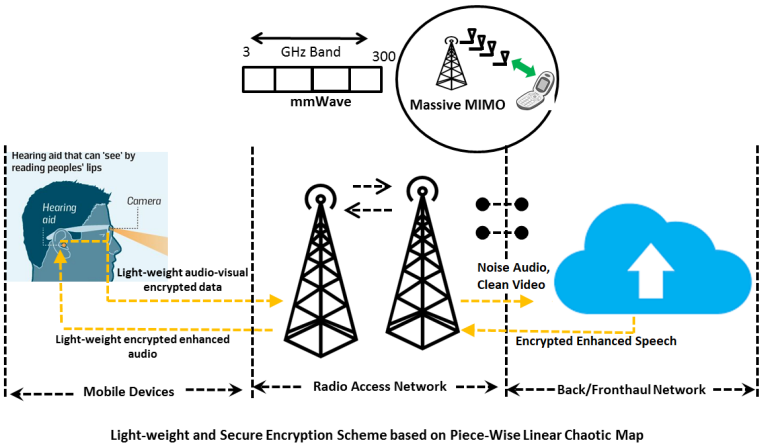


Figure 2: Proposed 5G IoT enabled Audio-Visual Hearing-Aid Framework

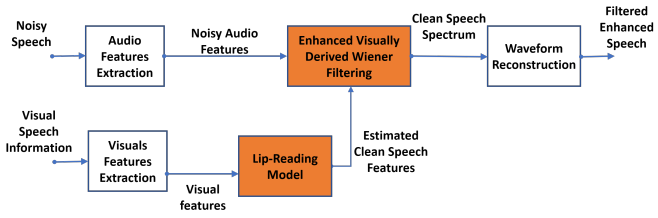


Figure 3: Proposed Lip-Reading Driven Deep Learning Approach for Speech Enhancement

The proposed 5G IoT enabled AV hearing-aid framework is depicted in Figure 2. It is to be noted that the computational complexity and real-time processing issues due to deep learning and big data optimization algorithms are addressed by running them in the background on the cloud. The mobile hearing-aid only transmits the encrypted compressed AV information and receives encrypted enhanced reconstructed speech in real-time. For end-to-end communication, an indoor 5G wireless small sized cell and back/fronthaul core network are proposed as the communication infrastructure of the system [29]. For IoT gateway, we propose the use of an efficient IoT gateway over a 5G wireless system, developed and tested in [16]. Specifically, we propose the use of an efficient IoT gateway over 5G wireless, which exploits small cells (with only 200 m radius), aggressive modulation and coding schemes (MCSs), massive MIMO, and high-frequency mmWave band (ranging from 3 to 300 GHz). It is to be noted that a small fraction of available mmWave spectrum is capable of supporting 100x more data rate and user capacity as compared to the state-of-the-art cellular spectrum [7, 8]. The proposed use of 5G CRAN with existing cloud computing services has the capacity to support thousands of connected devices in real-time.

The developed IoT gateway in [16] promises an uplink latency of 10ms and 5ms with and without compression respectively. In addition, it ensures minimum interference between medical IoT devices with low power consumption. The novelty of these gateways lie in efficient uplink IoT traffic classification and optimal uplink data (traffic) compression strategies. This helps in relaxation of the uplink traffic burden and results in efficient utilization of uplink wireless resources. More details on 5G-CRAN and front/backhaul connectivity are comprehensively presented in [16]. Ongoing and future work includes software integration of the proposed AV mobile hearing aid with 5G-CRAN and cloud computing as well as its hardware prototype implementation for real-time testing.

For real-time lightweight audio-visual encryption, piece-wise linear chaotic map (PWLCM), Chebyshev map, secure hash algorithm and novel S-Box algorithms are utilized. In the literature, conventional encryption approaches such as advanced encryption standard (AES) and Rivest–Shamir–Adleman (RSA)/Elliptic Curve (signing) are suitable for high processing power systems but incompatible with embedded low power sensor networks. Therefore, lightweight cryptography can potentially address real-time encryption challenges [30]. In our proposed scheme, the encrypted audio and video signals are exploited in the cloud by the designed novel lip-reading driven speech enhancement system, depicted in Figure 3. The proposed speech enhancement approach leverages the complementary strengths of both deep learning and analytical acoustic modelling (filtering based approach) that operates at two levels. In the first level, a novel deep learning based lip-reading regression model is employed. In the second level, lip-reading approximated clean-audio features are exploited, using an EVWF, for estimating the clean audio power spectrum. Finally, the Wiener filter is applied to the magnitude spectrum of the noisy input audio signal, followed by the inverse fast Fourier transform (IFFT), overlap, and combining processes to produce an enhanced magnitude spectrum. More details are presented in Section 4. The proposed AV speech enhancement framework finally transmits enhanced encrypted speech to the mobile hearing-aid.

3 Proposed Real-Time Lightweight Chaotic Encryption

In the proposed scheme, PWLCM, Chebyshev map, SHA and novel S-Box algorithms are effectively used for real-time lightweight encryption. The applied transformations are briefly explained in the subsequent sections.

3.1 Applied Transformations

3.1.1 PWLCM

As outlined in Shannon's novel paper [31], a good encryption scheme is composed of two stages: (i) Confusion and (ii) Diffusion. In the confusion stage, a correlation between key and ciphertext is made complex. Diffusion means

that a minor change in plaintext should change the corresponding ciphertext significantly. The proposed algorithm uses PWLCM in the confusion process. The PWLCM can be written as:

$$y_{n+1} = f(y_n, \lambda) = \begin{cases} \frac{y_n}{\lambda}, & \text{if } y_n \in [0, \lambda] \\ \frac{1-y_n}{1-\lambda}, & \text{if } y_n \in (\lambda, 0.5] \\ F(1-y_n), & \text{if } y_n \in (0.5, 1], \end{cases} \tag{1}$$

where, y_n are pseudo-random chaotic values, $y_n \in (0, 1)$, and λ is the control parameter. Both λ and y_0 serve as an initial condition and called as key for chaotic pseudo-random number generation.

3.1.2 Chebyshev Map

In [32], Huang et al., proposed a novel key generator method using the Chebyshev map. The Chebyshev map can be defined mathematically as [32, 33]:

$$T_k(x) = cos(k \times arc\ cos(x)), \tag{2}$$

where $k = 0, 1, 2, ...,N$ and $x \in [-1, 1]$. Huang suggested $k = 4$ for less computation and better use of Chebyshev which has been used in the proposed scheme, the Chebyshev function is given as:

$$f(x_i) = 8x_{i-1}^4 - 8x_{i-1}^2 + 1, i = 1, 2, ...N \tag{3}$$

3.1.3 Logistic-Sine Map

To overcome the drawbacks of a one-dimensional (1D) Logistic map, Zhou et al., in [34] proposed a novel method of chaotic maps combination. For a larger chaotic map, the authors combined two exiting 1D Logistic and Sine maps. A logistic-sine map is mathematically defined as [34]:

$$z_{n+1} = (rz_n(1 - z_n) + (4 - r)\frac{sin(\pi z_n)}{4})mod(1), \tag{4}$$

3.1.4 Secure Hash Algorithm (SHA)

SHA generates a fixed length value known as a hash code, by applying some function to the plaintext message. In the literature, SHA has different variants depending on the size of the output e.g, SHA-1, SHA-256 and SHA-512 for 128, 256, and 512 bits outputs, respectively. In the proposed scheme, we used SHA-512 such that $H(m) = h(512$ bits). The Secret Key in the proposed scheme is dependent on SHA-512. A minor change in plaintext generates a completely different hash and different initial key parameters.

3.1.5 Affine Transformation

Affine transformation is a one to one mapping that transforms a unique plaintext into a unique symbol. The following affine transformation is used in the proposed scheme:

$$AT(w) = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} w_7 \\ w_6 \\ w_5 \\ w_4 \\ w_3 \\ w_2 \\ w_1 \\ w_0 \end{bmatrix} \oplus \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \tag{5}$$

where w_i are coefficients of w i.e multiplicative inverse modulo $(w^8 + w^4 + w^2 + w^1 + 1)$.

3.2 Image Encryption Scheme

The flow chart for the grayscale images using PWLCM, Chebyshev, SHA-512 and affine transformation is shown in Figure 4. The detailed steps of our proposed cryptosystem are as follows:

- **Step 1**: Convert colour image I_c of size $A \times B$ to gray-scale image I_g and save result in ψ.
- **Step 2**: Apply SHA-512 on gray-scale plaintext image ψ and save hexadecimal hash value in variable θ.
- **Step 3**: Select first and last 12 hash values and save in κ_1 and κ_2.
- **Step 4**: Convert hexadecimal values saved in κ_1 and α_1 to decimal values and store result in κ_1, and β_2, respectively.
- **Step 5**: Generate SHA-based initial conditions for PWLCM and Chebyshev using following equations:

$$y_0 = \frac{\kappa_1}{2^{48}} \tag{6}$$

$$x_0 = \frac{\kappa_2}{2^{48}} \tag{7}$$

- **Step 6**: Iterate PWLCM A times and store chaotic values in α. Randomly permute rows of gray-scale image I_g using the sequence α and save values in I_{rp}.
- **Step 7**: Iterate Chebyshev map B times and store chaotic values in β. Randomly permute columns of I_{rp} using the sequence α and save values in $I_{permuted}$

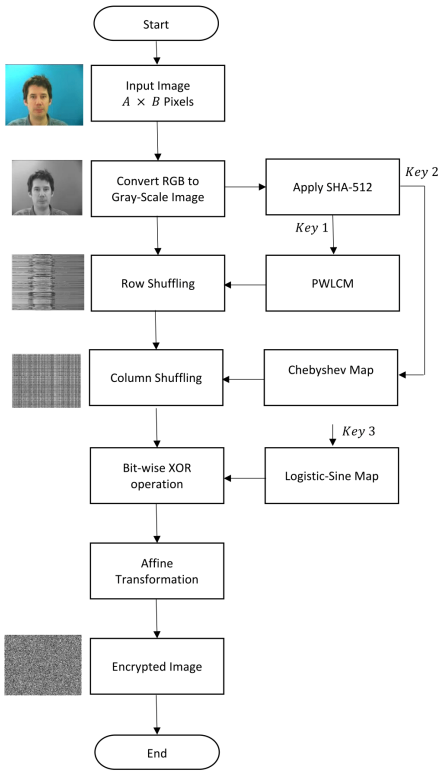


Figure 4: Proposed image encryption algorithm.

- **Step 8**: Iterate Logistic-sine map $A \times B$ times and store random values in γ.
- **Step 9**: Apply following operations on γ:

$$R_1 = \text{Mod}(\gamma \times 10^{14}, 256), \tag{8}$$

$$R_2 = floor(R_1). \tag{9}$$

- **Step 10**: Rearrange row-vector R_2 in matrix form R and Bit-wise XOR random matrix R with $I_{permuted}$ to get ϕ.
- **Step 11**: Apply affine transformation on ϕ and store values as a ciphertext image C.

For decryption, encryption steps are followed in reverse order.

3.3 Audio Encryption Scheme

The flow chart for the audio signal using PWLCM, Chebyshev, SHA-512 and affine transformation is shown in Figure 5. The detailed steps of our proposed cryptosystem are as follows:

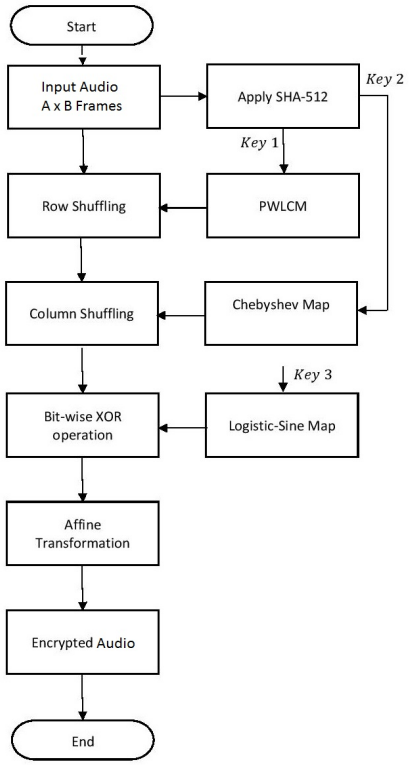


Figure 5: Proposed audio encryption algorithm.

- **Step 1**: Convert 1D audio signal into 2D of size $A \times B$ and save result in ψ.
- **Step 2**: Apply SHA-512 on ψ and save hexadecimal hash value in variable θ.
- **Step 3**: Select first and last 12 hash values and save in κ_1 and κ_2.
- **Step 4**: Convert hexadecimal values saved in κ_1 and α_1 to decimal values and store results in κ_1, and β_2, respectively.
- **Step 5**: Generate SHA-based initial conditions for PWLCM and Chebyshev using equations (6-7)

- **Step 6**: Iterate PWLCM A times and store chaotic values in α. Randomly permute rows of audio vector A_g using the sequence α and save values in I_{rp}.
- **Step 7**: Iterate Chebyshev map B times and store chaotic values in β. Randomly permute columns of A_{rp} using the sequence α and save values in $A_{permuted}$
- **Step 8**: Iterate Logistic-sine map $A \times B$ times and store random values in γ.
- **Step 9**: Apply operations given in equations (8-9) to γ.
- **Step 10**: Rearrange row-vector R_2 in matrix form R and Bit-wise XOR random matrix R with $A_{permuted}$ to get ϕ.
- **Step 11**: Apply affine transformation on ϕ and store values as a ciphertext audio C.

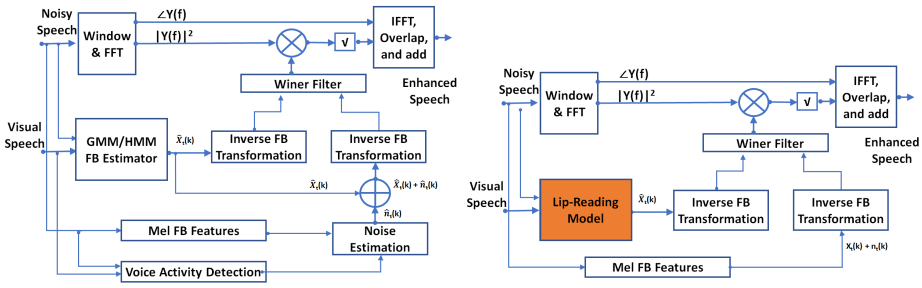
For decryption, encryption steps are followed in reverse order.

4 Speech Enhancement Framework

The state-of-the-art VWF and designed EVWF are depicted in Figure 6 (a) and (b) respectively. The authors in [15] presented a hidden Markov model-Gaussian mixture model (HMM/GMM) based two-level state-of-the-art VWF for speech enhancement. However, the use of HMM/GMM models for the estimation of clean audio features from visual features, and cubic spline interpolation for the approximation of high dimensional clean audio power spectrum from estimated low dimensional audio features, are not optimal choices. The HMM/GMM model suffers from poor generalization and the cubic spline interpolation method fails to estimate the missing power spectral values that lead to poor audio power spectrum estimation. In contrast, the designed EVWF addressed the limitations of the state-of-the-art VWF [15] by employing an inverse filter-bank transformation (i.e. a pseudoinverse of the approximated audio features) for audio power spectrum estimation, as compared to the cubic spline interpolation method. In addition, the use of LSTM addressed the generalization and accurate clean speech coefficient estimation issues. The designed EVWF also eliminates the need for voice activity detection (VAD) and noise estimation. More details are comprehensively presented in our previous work [23]

4.1 Lip Reading Model

The designed LSTM based lip-reading model consists of an input layer, two LSTM layers, and an output dense layer. In the designed LSTM model, prior visual features were fed into the stacked LSTM layers to exploit the existing temporal correlation. The lower LSTM layer used 250 cells for encoding the input



(a) State-of-the-art visually-derived Wiener filtering [15].

(b) Proposed enhanced visually-derived Wiener filtering

Figure 6: State-of-the-art visually-derived Wiener filtering (a) and proposed enhanced visually-derived Wiener filtering (b)

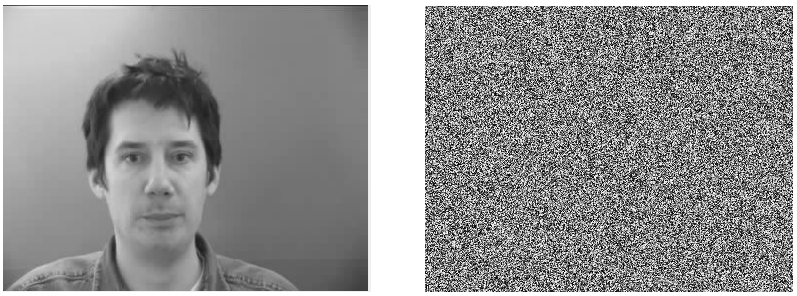
visual information and passed its hidden state to the second LSTM layer, which has 300 cells. The output of the second LSTM layer was then fed into the fully connected (dense) layer which has a total of 23 neurons with linear activation function. The designed LSTM model was trained with an objective to minimise the mean squared error (MSE) between the predicted and the actual audio features, using a stochastic gradient decent algorithm and RMSProp optimiser. More dataset, pre-processing, and training/testing details are comprehensively presented in our previous works [22][23].

5 Dataset and Audio-Visual Feature Extraction

For AV encryption and speech enhancement, Grid [35] and ChiME3 [36] corpora are used. The proposed system is evaluated under four different dynamic real-world scenarios (cafe, street, public transport, pedestrian area). It is to be noted that the utterances from all the scenarios were mixed to develop a contextual AV speech enhancement framework. The visual only speech enhancement significantly outperforms audio-only approaches at low SNRs. However, visual cues become less effective at high SNRs. Therefore, to effectively account for different noisy conditions, a more optimal, context-aware audio-visual system is required, that leverages the complementary strengths of both visual and noisy audio cues contextually. For noisy utterance generation, clean videos from Grid corpus were mixed with ChiME3 noises for different SNR levels ranging from -12 to 12dB. For preprocessing, sentence alignment and prior visual frames were used to stop the model from learning redundant information and improve mapping between visual and audio features (exploiting their temporal information).

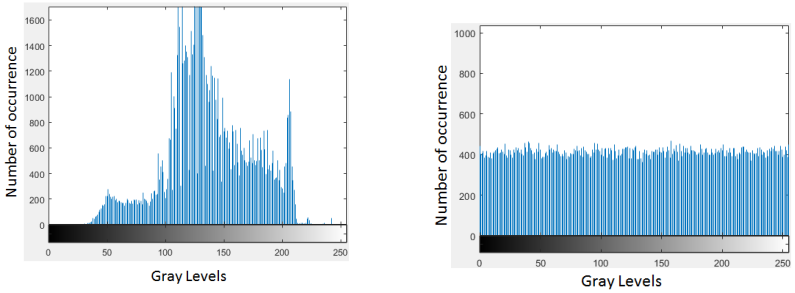
5.1 Audio-Visual feature extraction

For audio feature extraction, the input audio signal was sampled at 50kHz and segmented into N 16ms frames with 800 samples per frame and a 62.5% increment rate. To produce a 2048-bin power spectrum, a hamming window



(a) Original Image (b) Encrypted Image

Figure 7: Encryption Results. It is to be noted that the proposed encryption scheme completely concealed the plaintext information.



(a) Original Image Histogram. (b) Encrypted Image Histogram.

Figure 8: Histogram Results. It is to be noted that the obtained histogram of the encrypted image is flat which is ideally required.

and Fourier transformation was applied, followed by a logarithmic compression to produce a 23-D log-FB signal. The visual features were extracted from the Grid Corpus videos recorded at 25 fps. The video files were processed by extracting a sequence of individual frames and applying a Viola-Jones lip detector [37] and object tracker [38]. Furthermore, to ensure appropriate lip tracking, processed utterances were manually validated. Finally, the 2D-Discrete Cosine Transformation (2D-DCT) was applied to produce vectors of pixel intensities, followed by interpolation. More details are presented in our previous work [22].

6 Performance Evaluation

6.1 Lip-Images Encryption and Security Analyses

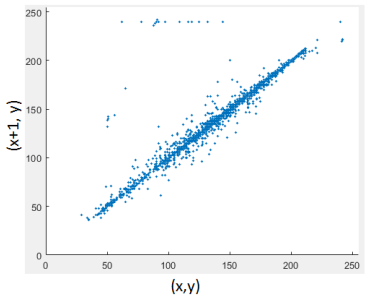
In order to show the effectiveness of the encryption scheme, one test image is selected with a particular lips position. The proposed lightweight encryption scheme is applied to the plaintext image and the obtained results are shown in Figure 7. It can be seen that the proposed scheme completely concealed

the plaintext information. Moreover, the histogram results shown in Figure 8 acquired the flat histogram which is required ideally. However, the results are not sufficient to prove complete security of the cryptosystem. Therefore, a large number of security metrics such as correlation coefficient, entropy, contrast, energy, number of pixel change rate (NPCR), and unified average change intensity (UACI) defined in our previous work [39, 40, 41] are used. The degree of similarity between adjacent pixels are generally analysed via correlation coefficient metrics. Ideally, correlation in all directions (horizontal (H_{CC}), vertical(V_{CC}) and diagonal (D_{CC})) should be close to zero. Entropy is another important metric which can evaluate resistance capability against statistical attacks. For a good cryptosystem, entropy of a gray-scale encrypted image should be 8 bits. A contrast of an encrypted image is defined as the intensity between a pixel and its neighbour pixels. In image encryption, higher values of contrast indicate a higher quality of encrypted image. A sum of squared elements (SSE) in a gray level co-occurrence matrix returns the energy of an image. The energy value of a secure encrypted image is desired to be low. In case of complete constant pixels, the energy value is 1. NPCR and UACI show resistance against differential attacks. Higher values of NPCR and UACI reflect higher encryption quality. More details on how these parameters prove the security of our encryption scheme is defined in detail in our previous work [39, 40, 41].
The correlation plots in vertical directions are shown in Figure 9. From these, it is evident that the distribution of adjacent pixels in a vertical direction is uncorrelated as compared to plaintext correlation. Mathematical values of correlations in vertical, horizontal and diagonal directions are outlined in Table 1. In the table, lower correlation values prove the robustness of the proposed scheme. In addition, it can be seen that the histogram plot for the proposed scheme is uniformly distributed; hence, assures resistance to statistical attacks as compared to existing algorithms [42][43]. NPCR greater than 99% also reveals higher security. Consequently, the security metrics demonstrate the effectiveness and higher security of the proposed scheme. Lastly, the required encryption/decryption time is less than 25 msec on a 60 GB RAM using MATLAB software. Such a low processing time confirms that the proposed scheme is lightweight and could be an effective solution for practical real-time applications. However, further real-time optimization is an ongoing work.

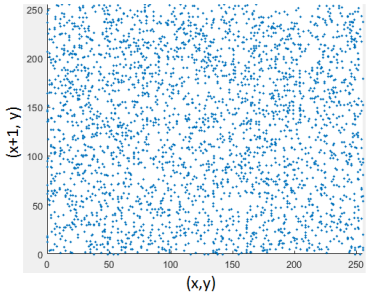
6.2 Speech Enhancement Results

6.3 Lip-Reading Results

For lip-reading, multiple prior visual frames (lip images) are used (ranging from 1 visual frame to 27 prior visual frames). The simulation results are shown in Table 2. It can be seen that by moving from 1 visual frame to 18 visual frames, a significant performance improvement could be achieved. The LSTM model with 1 visual frame achieved the MSE of 0.092, whereas with 18 visual frames the model achieved the least MSE of 0.058. LSTM based learning model exploited the temporal information (i.e. prior visual frames) efficiently and



(a) Original Image correlation coefficient.



(b) Encrypted Image correlation coefficient.

Figure 9: Correlation coefficients in vertical direction.

Table 1: Image security assessment. It is to be noted that the correlation values in all directions (horizontal (H_{CC}), vertical(V_{CC}) and diagonal (D_{CC})) are low which proves the robustness of the proposed scheme. In addition, note the resistance capability against statistical attacks and high security, evaluated using Contrast, Energy, NPCR and UACI tests.

Security Parameter	Original Frame	Encrypted Frame
V_{CC}	0.9523	0.0044
H_{CC}	0.9711	-0.0056
D_{CC}	0.9677	0.0089
$Entropy$	7.0025	7.9983
$Contrast$	0.1049	10.4608
$Energy$	0.2161	0.0156
$NPCR$	NA	99.4566
$UACI$	NA	33.1561

Table 2: LSTM Training and Testing Accuracy - Comparison For Different Visual Frames. The table presents an overall behaviour of the LSTM model when contextual information (i.e. previous frames) is added. It is to be noted that the LSTM model exploited the temporal correlation effectively but saturated at 18 prior visual frames.

	LSTM	
Visual Frames	MSE_{train}	MSE_{test}
1	0.092	0.104
2	0.087	0.097
4	0.073	0.085
8	0.066	0.082
14	0.061	0.080
18	0.058	0.078

Table 3: Speech Enhancement Results. It can be seen that at low SNR levels, EVWF significantly outperformed benchmark SS and LMMSE based speech enhancement methods.

SNR	**PESQ**			**MOS**		
	SS	**LMMSE**	**EVWF**	**SS**	**LMMSE**	**EVWF**
-12dB	0.9	0.95	1.52	0.31	0.315	1.68
-6dB	1.01	1.03	1.58	0.51	0.48	1.965
-3dB	1.17	1.18	1.60	1.17	1.165	2
0dB	1.21	1.20	1.63	1.925	1.97	2.11
3dB	1.25	1.34	1.72	2.025	2.085	2.22
6dB	1.26	1.39	1.69	2.295	2.345	2.33
12dB	1.54	1.60	1.74	2.58	2.61	2.54

showed consistent reduction in MSE while going from 1 to 18 visual frames. This is mainly because of its inherent recurrent architectural property and the ability of retaining state over long time spans by using gates. More details, critical analysis, and comparisons are comprehensively presented in our previous work [22].

6.3.1 Objective Test

For objective testing and comparison with state-of-the-art audio only speech enhancement methods (spectral subtraction (SS) and Log-Minimum Mean Square Error (LMMSE)), perceptual evaluation of speech quality (PESQ) is used to evaluate the quality of restored speech. PESQ is one of the most reliable methods to evaluate speech quality. The PESQ score is computed as a linear combination of the average disturbance value and the average asymmetrical disturbance values. Scores range from -0.5 to 4.5, corresponding with low to high speech quality. The PESQ scores for the proposed EVWF and state-of-the-art benchmark audio only speech enchantment approaches are depicted in Table 3. It can be seen that at low SNR levels, EVWF significantly outperformed both SS and LMMSE

Table 4: Audio security assessment.

Security Parameter	Encrypted Signal
NSCR	99.95%
UACI	33.39%
Correlation coefficient	0.00022
Key length	10^{45}

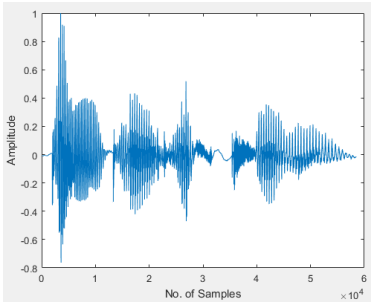
based speech enhancement methods.

6.3.2 Subjective Listening Tests

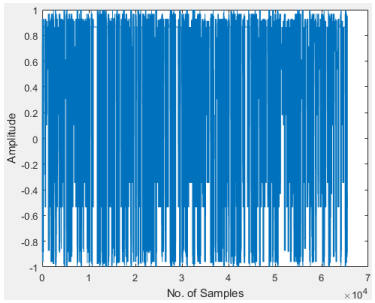
The subjective listening test was conducted in terms of MOS with self-reported normal-hearing listeners. The listeners were presented with both clean (target) and enhanced speech, and were asked to rate the re-constructed speech on a scale of 1 to 5. The five rating choices were: (5) Excellent (when the listener feels an unnoticeable difference compared to the target clean speech) (4) Good (perceptible but not annoying) (3) Fair (slightly annoying) (2) Poor (annoying), and (1) Bad (very annoying). The EVWF performance was compared with two state-of-the-art speech enhancement methods (SS and LMMSE). A total of 10 listeners took part in the evaluation session. In Table 3, it can be seen that at low SNRs (-12dB, -6dB, and -3dB), the proposed EVWF outperformed audio-only speech enhancement methods. On the other hand, for high SNRs, our AV approach performed comparably to the Audio only approach.

6.4 Audio Encryption and Security Analyses

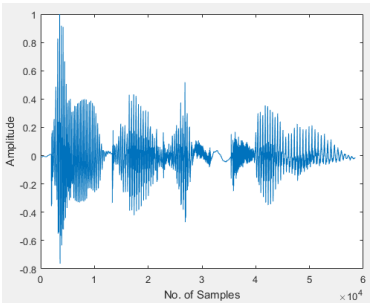
The proposed lightweight chaotic encryption was applied to the enhanced audio signal. A single bit in the original speech signal was modified and encrypted via the same key [44]. The two ciphered speech signals, i.e., S_c1 and S_c2 were generated and the obtained results are shown in Figure 10. It can be seen that the proposed scheme completely concealed the audio information and decrypted speech signal accurately. Furthermore, to check the robustness and security strength of the proposed audio encryption scheme, security analysis tests such as number of sample change rate (NSCR), UACI, correlation coefficient, and key length were conducted, where low correlation values, 99.95% NSCR, and 33.3% UACI demonstrate the robustness of the proposed encryption scheme. In addition, the key length of the proposed approach is much greater than the minimum required length (2^{100}) which shows resistance against brute force attack. Table 4 presents the security analysis and effectiveness of the proposed scheme.



(a) Original Audio



(b) Encrypted Audio



(c) Decrypted Audio

Figure 10: Audio encryption results.

7 Conclusions and future directions

Next-generation multimodal hearing-aids stand as a major enabler for modern digital hearing aids, capable of restoring intelligibility and reducing cognitive load in environments with overwhelming noise. However, the real-time implementation of such AV hearing-aids demand high data rate, low latency, low computational complexity, and high security. In this paper, we proposed a 5G IoT enabled AV hearing-aid framework that leverages the complementary strengths of 5G and IoT technology to address the aforementioned challenges. As part of the envisioned technology, two main contributions (AV speech enhancement in the cloud and lightweight AV encryption) are reported here. Comparative critical analysis in terms of both speech enhancement and AV encryption demonstrate the potential of the envisioned technology to deliver high quality speech reconstruction in extreme noisy situations, and secure mobile AV hearing aid communication. Specifically, the comparative performance evaluation of the proposed speech enhancement method under real noisy environments revealed that the proposed approach significantly outperformed benchmark audio-only approaches at low SNR, with comparable performances at high SNRs. The audio and video encryption results revealed the effectiveness of the proposed real-time lightweight encryption scheme in terms of both high security and processing time. The ongoing/future work includes utilizing partial encryption [39] to further reduce the encryption and decryption time, whilst maintaining required data security. The partial encryption will help address high power consumption or quick battery drainage issues. Further work will include software integration of the proposed AV mobile hearing aid with 5G-CRAN and its hardware prototype implementation for real-time prototyping and testing. Market assessment will also be carried out, where the aim is to identify potential opportunities in the UK hearing industry.

Acknowledgments

scThe authors would like to gratefully acknowledge Mandar Gogate from the Edinburgh Napier University for his contribution in implementing LSTM driven AV mapping, which was published in our previous work and cited here for reference.

Funding

This research was supported by Engineering and Physical Sciences Research Council (EPSRC) Grant No. EP/M026981/1 and deepCI grant No.DCI1012.

Compliance with Ethical Standards

This manuscript has not been published in whole or in part elsewhere, which has also not currently being considered for publication in another journal. All

authors have been personally and actively involved in substantive work leading to the manuscript, and will hold themselves jointly and individually responsible for its content.

Conflict of Interest

The authors declare that they have no conflict of interest.

Ethical Approval

This article does not contain any studies with human participants performed by any of the authors.

Author's Contributions

AA and AH conceived and developed the original idea reported in this paper, of integrating 5G, IoT, and lightweight encryption, with the lip-reading driven hearing-aid. AA and JA performed the simulations.

References

[1] Shibli Nisar, Muhammad Tariq, Ahsan Adeel, Mandar Gogate, and Amir Hussain. Cognitively inspired feature extraction and speech recognition for automated hearing loss testing. *Cognitive Computation*, pages 1–14, 2019.

[2] Hearing Aids Market. `https://www.marketsandmarkets.com/PressReleases/hearing-aids.asp`. Accessed: 2019-02-15.

[3] AJ Ruggles and IW Ekoto. Ignitability and mixing of underexpanded hydrogen jets. *International Journal of Hydrogen Energy*, 37(22):17549–17560, 2012.

[4] S Kortlang, S Ewert, H Meister, S Rählmann, J Kießling, et al. Combination of controlled laboratory tests and structured field trials for a comprehensive evaluation of a model-based hearing aid. *Int J Audiol*, 2016.

[5] Rudy Rotili, Emanuele Principi, Stefano Squartini, and Björn Schuller. A real-time speech enhancement framework in noisy and reverberated acoustic scenarios. *Cognitive Computation*, 5(4):504–516, 2013.

[6] Joyner Cadore, Francisco J Valverde-Albacete, Ascensión Gallardo-Antolín, and Carmen Peláez-Moreno. Auditory-inspired morphological processing of speech spectrograms: Applications in automatic speech recognition and speech enhancement. *Cognitive computation*, 5(4):426–441, 2013.

[7] MA Ben Messaoud, Aïcha Bouzid, and Noureddine Ellouze. A new biologically inspired fuzzy expert system-based voiced/unvoiced decision algorithm for speech enhancement. *Cognitive Computation*, 8(3):478–493, 2016.

[8] Ravi Kumar Kandagatla and PV Subbaiah. Speech enhancement using mmse estimation of amplitude and complex speech spectral coefficients under phase-uncertainty. *Speech Communication*, 96:10–27, 2018.

[9] Ali I Siam, Heba A El-khobby, Mustafa M Abd Elnaby, Hatem S Abdelkader, and Fathi E Abd El-Samie. A novel speech enhancement method using fourier series decomposition and spectral subtraction for robust speaker identification. *Wireless Personal Communications*, pages 1–14, 2019.

[10] Amir Hussain, Jon Barker, Ricard Marxer, Ahsan Adeel, William Whitmer, Roger Watt, and Peter Derleth. Towards multi-modal hearing aid design and evaluation in realistic audio-visual settings: Challenges and opportunities. *First International Conference on Challenges in Hearing assistive Technology (CHAT-17) Stockholm, Sweden, August 19.2017*, 2017.

[11] William H Sumby and Irwin Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, 1954.

[12] Quentin Summerfield. Use of visual information for phonetic perception. *Phonetica*, 36(4-5):314–331, 1979.

[13] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746, 1976.

[14] Michelle L Patterson and Janet F Werker. Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, 6(2):191–196, 2003.

[15] Ben Almajai, Milner. Visually derived wiener filters for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1642–1651, 2011.

[16] Navrati Saxena, Abhishek Roy, Bharat JR Sahu, and HanSeok Kim. Efficient iot gateway over 5g wireless: A new design with prototype and implementation results. *IEEE Communications Magazine*, 55(2):97–105, 2017.

[17] Fadi Al-Turjman, Enver Ever, and Hadi Zahmatkesh. Small cells in the forthcoming 5g/iot: Traffic modelling and deployment overview. *IEEE Communications Surveys & Tutorials*, 2018.

[18] Fadi Al-Turjman. Fog-based caching in software-defined information-centric networks. *Computers & Electrical Engineering*, 69:54–67, 2018.

[19] Mohammed Zaki Hasan, Fadi Al-Turjman, and Hussain Al-Rizzo. Analysis of cross-layer design of quality-of-service forward geographic wireless sensor network routing strategies in green internet of things. *IEEE Access*, 6:20371–20389, 2018.

[20] Fadi Al-Turjman. Cognitive caching for the future sensors in fog networking. *Pervasive and Mobile Computing*, 42:317–334, 2017.

[21] Fadi Al-Turjman and Sinem Alturjman. Confidential smart-sensing framework in the iot era. *The Journal of Supercomputing*, 74(10):5187–5198, 2018.

[22] Ahsan Adeel, Mandar Gogate, Amir Hussain, and William M Whitmer. Lip-reading driven deep learning approach for speech enhancement. *IEEE Transactions on Emerging Topics in Computational Intelligence (in-press)*, 2019.

[23] Ahsan Adeel, Mandar Gogate, and Amir Hussain. Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments. *Information Fusion (In Press)*, 2019.

[24] Ahsan Adeel, Hadi Larijani, and Ali Ahmadinia. Random neural network based novel decision making framework for optimized and autonomous power control in lte uplink system. *Physical Communication*, 19:106–117, 2016.

[25] Richard Einhorn. Hearing aid technology for the 21st century: A proposal for universal wireless connectivity and improved sound quality. *IEEE pulse*, 8(2):25–28, 2017.

[26] Mamta Agiwal, Abhishek Roy, and Navrati Saxena. Next generation 5g wireless networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 18(3):1617–1655, 2016.

[27] Jeffrey G Andrews, Stefano Buzzi, Wan Choi, Stephen V Hanly, Angel Lozano, Anthony CK Soong, and Jianzhong Charlie Zhang. What will 5g be? *IEEE Journal on selected areas in communications*, 32(6):1065–1082, 2014.

[28] Naga Bhushan, Junyi Li, Durga Malladi, Rob Gilmore, Dean Brenner, Aleksandar Damnjanovic, Ravi Sukhavasi, Chirag Patel, and Stefan Geirhofer. Network densification: the dominant theme for wireless evolution into 5g. *IEEE Communications Magazine*, 52(2):82–89, 2014.

[29] Min Chen, Jun Yang, Yixue Hao, Shiwen Mao, and Kai Hwang. A 5g cognitive system for healthcare. *Big Data and Cognitive Computing*, 1(1):2, 2017.

[30] William J Buchanan, Shancang Li, and Rameez Asif. Lightweight cryptography methods. *Journal of Cyber Security Technology*, 1(3-4):187–201, 2017.

[31] Claude E Shannon. Communication theory of secrecy systems. *Bell Labs Technical Journal*, 28(4):656–715, 1949.

[32] Xiaoling Huang. Image encryption algorithm using chaotic chebyshev generator. *Nonlinear Dynamics*, 67(4):2411–2417, 2012.

[33] Xingyuan Wang, Dapeng Luan, and Xuemei Bao. Cryptanalysis of an image encryption algorithm using chebyshev generator. *Digital Signal Processing*, 25:244–247, 2014.

[34] Yicong Zhou, Long Bao, and CL Philip Chen. A new 1d chaotic system for image encryption. *Signal processing*, 97:172–182, 2014.

[35] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[36] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. The third 'chime'speech separation and recognition challenge: Dataset, task and baselines. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 504–511. IEEE, 2015.

[37] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001.

[38] David A Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 77(1-3):125–141, 2008.

[39] Jawad Ahmad, Muazzam A Khan, Seong Oun Hwang, and Jan Sher Khan. A compression sensing and noise-tolerant image encryption scheme based on chaotic maps and orthogonal matrices. *Neural computing and applications*, 28(1):953–967, 2017.

[40] Fadia Ali Khan, Jameel Ahmed, Jan Sher Khan, Jawad Ahmad, and Muazzam A Khan. A novel substitution box for encryption based on lorenz equations. In *Circuits, System and Simulation (ICCSS), 2017 International Conference on*, pages 32–36. IEEE, 2017.

[41] Jan Sher Khan, Jawad Ahmad, and Muazzam A Khan. Td-ercs map-based confusion and diffusion of autocorrelated data. *Nonlinear Dynamics*, 87(1):93–107, 2017.

[42] Jawad Ahmad and Seong Oun Hwang. Chaos-based diffusion for highly autocorrelated data in encryption algorithms. *Nonlinear Dynamics*, 82(4):1839–1850, 2015.

[43] Amir Anees, Adil Masood Siddiqui, and Fawad Ahmed. Chaotic substitution for highly autocorrelated data in encryption algorithm. *Communications in Nonlinear Science and Numerical Simulation*, 19(9):3106–3118, 2014.

[44] P Sathiyamurthi and S Ramakrishnan. Speech encryption using chaotic shift keying for secured speech communication. *EURASIP Journal on Audio, Speech, and Music Processing*, 2017(1):20, 2017.